Part 3

1. The GPT generated test pairs clearly reflected a causal relationship (i.e., only one protected attribute changed while others stayed the same)

○ Strongly disagree

○ Disagree

○ Neutral

○ Agree

○ Strongly agree

2. The GPT generated tests helped me understand how discrimination could arise in software behavior.

○ Strongly disagree

○ Disagree

○ Neutral

○ Agree

○ Strongly agree

3. The GPT-generated input test cases covered a wide and representative range of input combinations (including variations in both protected and non-protected attributes).

○ Strongly disagree – The generated tests were narrow and repetitive

○ Disagree – The tests explored limited attribute combinations

○ Neutral – The coverage was moderate but could be improved

○ Agree – The tests showed good diversity across attributes

○ Strongly agree – The tests comprehensively covered the input space with diverse and meaningful variations

4. Which tool generated clearer and more logically consistent causal test pairs (where only one protected attribute changed)?

○ Strongly prefer Themis

○ Slightly prefer Themis

○ No preference

○ Slightly prefer GPT

○ Strongly prefer GPT

5. Which tool made you feel having more control over the discrimination testing space??

○ Strongly prefer Themis

○ Slightly prefer Themis

○ No preference

○ Slightly prefer GPT

○ Strongly prefer GPT

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms